1. Detection Theory Basics
2. Detection Performance Criteria
3. Detection of Signals in Noise
4. Beyond the Basics

Detection Theory Basics

Detection theory concerns making decisions from data. Decisions are based on presumptive models that may have produced the data. Making a decision involves inspecting the data and determining which model was most likely to have produced them. In this way, we are detecting which model was correct. Decision problems pervade signal processing. In digital communications, for example, determining if the current bit received in the presence of channel disturbances was a zero or a one is a detection problem.

More concretely, we denote by $\mathscr{M}_{\mathrm{i}}$ the i^{th} model that could have generated the data $\boldsymbol{R}$. A "model" is captured by the conditional probability distribution of the data, which is denoted by the vector $\boldsymbol{R}$. For example, model i is described by $\mathrm{p}_{\boldsymbol{R}|\mathscr{M}_i}\left(\boldsymbol{r}\right)$. Given all the models that can describe the data, we need to choose which model best matched what was observed. The word "best" is key here: what is the optimality criterion, and does the detection processing and the decision rule depend heavily on the criterion used? Surprisingly, the answer to the second question is "No." All of detection theory revolves around the **likelihood ratio test**, which as we shall see, emerges as the optimal detector under a wide variety of optimality criteria.

The Likelihood Ratio Test

In a binary detection problem in which we have two models, four possible decision outcomes can result. Model $\mathscr{M}_0$ did in fact represent the best model for the data and the decision rule said it was (a correct decision) or said it wasn't (an erroneous decision). The other two outcomes arise when model $\mathscr{M}_1$ was in fact true with either a correct or incorrect decision made. The decision process operates by segmenting the range of observation values into two disjoint **decision regions** Z_0 and Z_1. All values of $\boldsymbol{R}$ fall into either Z_0 or Z_1. If a given $\boldsymbol{R}$ lies in Z_0, we will announce our decision ""model $\mathscr{M}_0$ was true""; if in Z_1, model $\mathscr{M}_1$ would be proclaimed. To derive a rational method of deciding which model best describes the observations, we need a criterion to assess the quality of the decision process so that optimizing this criterion will specify the decision regions.

The **Bayes' decision criterion** seeks to minimize a cost function associated with making a decision. Let C_{ij} be the cost of mistaking model j for model i ($i \neq j$) and C_{ii} the presumably smaller cost of correctly choosing model i: $C_{ij} > C_{ii}$, $i \neq j$. Let π_i be the a priori probability of model i. The so-called **Bayes'**

cost C is the average cost of making a decision.
Equation:

$$\begin{aligned} C &= \sum_{i,j\{i,j\}} C_{ij} \Pr[\mathrm{say}\mathscr{M}_i\mathrm{when}\mathscr{M}_j\mathrm{true}] \\ &= \sum_{i,j\{i,j\}} C_{ij}\pi_j \Pr[\mathrm{say}\mathscr{M}_i \mid \mathscr{M}_j\mathrm{true}] \end{aligned}$$

The Bayes' cost can be expressed as
Equation:

$$\begin{aligned} C &= \sum_{i,j\{i,j\}} C_{ij}\pi_j \Pr[\boldsymbol{R} \in Z_i \mid \mathscr{M}_{\mathrm{j}}\mathrm{true}] \\ &= \sum_{i,j\{i,j\}} C_{ij}\pi_j \int \mathrm{p}_{\boldsymbol{R}|\mathscr{M}_j}\left(\boldsymbol{r}\right) \mathrm{d}\,\boldsymbol{r} \\ &= \int \left(C_{00}\pi_0\, \mathrm{p}_{\boldsymbol{R}|\mathscr{M}_0}\left(\boldsymbol{r}\right) + C_{01}\pi_1\, \mathrm{p}_{\boldsymbol{R}|\mathscr{M}_1}\left(\boldsymbol{r}\right)\right) \mathrm{d}\,\boldsymbol{r} + \int \left(C_{10}\pi_0\, \mathrm{p}_{\boldsymbol{R}|\mathscr{M}_0}\left(\boldsymbol{r}\right) + C_{11}\pi_1\, \mathrm{p}_{\boldsymbol{R}|\mathscr{M}_1}\left(\boldsymbol{r}\right)\right) \mathrm{d}\,\boldsymbol{r} \end{aligned}$$

To minimize this expression with respect to the decision regions Z_0 and Z_1, ponder which integral would yield the smallest value if its integration domain included a specific value of the observation vector. To

minimize the sum of the two integrals, whichever integrand is smaller should include that value of $\boldsymbol{r}$ in its integration domain. We conclude that we choose $\mathscr{M}_0$ for those values of $\boldsymbol{r}$ yielding a smaller value for the first integral.

$$\pi_0 C_{00}\ \mathrm{p}_{\boldsymbol{R}|\mathscr{M}_0}(\boldsymbol{r}) + \pi_1 C_{01}\ \mathrm{p}_{\boldsymbol{R}|\mathscr{M}_1}(\boldsymbol{r}) < \pi_0 C_{10}\ \mathrm{p}_{\boldsymbol{R}|\mathscr{M}_0}(\boldsymbol{r}) + \pi_1 C_{11}\ \mathrm{p}_{\boldsymbol{R}|\mathscr{M}_1}(\boldsymbol{r})$$

We choose $\mathscr{M}_1$ when the inequality is reversed. This expression is easily manipulated to obtain the crowning result of detection theory: the **likelihood ratio test**.
Equation:

$$\frac{\mathrm{p}_{\boldsymbol{R}|\mathscr{M}_1}(\boldsymbol{r})}{\mathrm{p}_{\boldsymbol{R}|\mathscr{M}_0}(\boldsymbol{r})} \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \frac{\pi_0 (C_{10} - C_{00})}{\pi_1 (C_{01} - C_{11})}$$

The comparison relation means selecting model $\mathscr{M}_1$ if the left-hand ratio exceeds the value on the right; otherwise, $\mathscr{M}_0$ is selected. The **likelihood ratio** $\frac{\mathrm{p}_{\boldsymbol{R}|\mathscr{M}_1}(\boldsymbol{r})}{\mathrm{p}_{\boldsymbol{R}|\mathscr{M}_0}(\boldsymbol{r})}$, symbolically represented by $\Lambda(\boldsymbol{r})$, encapsulates the signal processing performed by the optimal detector on the observations $\boldsymbol{r}$. The optimal decision rule then compares that scalar-valued result with a **threshold** η equaling $\frac{\pi_0(C_{10}-C_{00})}{\pi_1(C_{01}-C_{11})}$. The likelihood ratio test can be succinctly expressed as the comparison of the likelihood ratio with a threshold.
Equation:

$$\Lambda(\boldsymbol{r}) \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \eta$$

The data processing operations are captured entirely by the likelihood ratio $\Lambda(\boldsymbol{r})$. However, the calculations required by the likelihood ratio can be simplified in many cases. Note that only the value of the likelihood ratio **relative** to the threshold matters. Consequently, we can perform **any** positively monotonic transformation simultaneously on the likelihood ratio and the threshold without affecting the result of the comparison. For example, we can multiply by a positive constant, add any constant or apply a monotonically increasing function to reduce the complexity of the expressions. We single out one such function, the logarithm, because it often simplifies likelihood ratios that commonly occur in signal processing applications. Known as the **log-likelihood**, we explicitly express the likelihood ratio test with it as
Equation:

$$\ln(\Lambda(\boldsymbol{r})) \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \ln(\eta)$$

What simplifying transformations are useful are problem-dependent. But, by laying bare what aspect of the observations is essential to the model-testing problem, we reveal the **sufficient statistic** $\Upsilon(\boldsymbol{r})$: the scalar quantity which best summarizes the data for detection purposes. The likelihood ratio test is best expressed in terms of the sufficient statistic.
Equation:

$$\Upsilon(\boldsymbol{r}) \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \gamma$$

We denote the threshold value for the sufficient statistic by γ or by η when the likelihood ratio is used in the comparison.

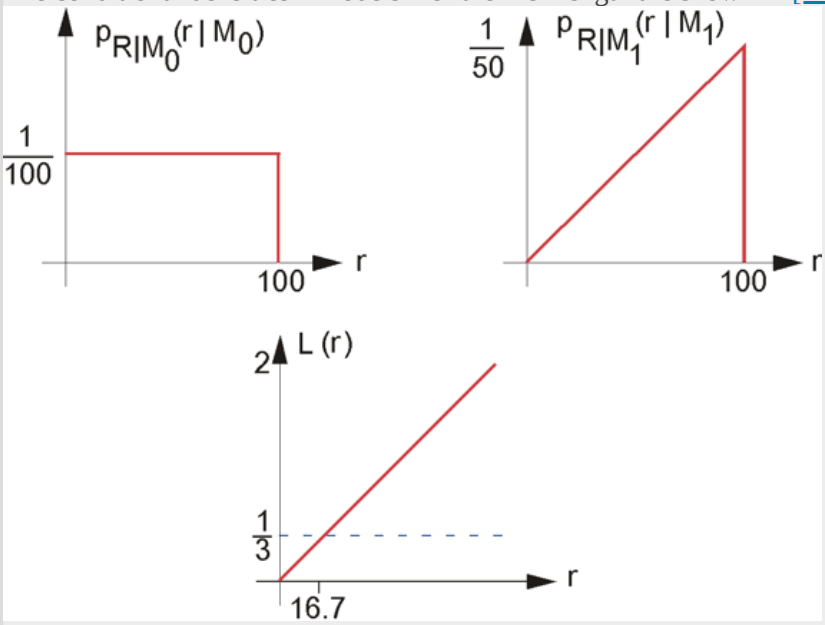
The likelihood ratio is comprised of the quantities $p_{\boldsymbol{R}|\mathcal{M}_i}(\boldsymbol{r})$, which are known as **likelihood functions** and play an important role in estimation theory. It is the likelihood function that portrays the probabilistic model describing data generation. The likelihood function completely characterizes the kind of "world" assumed by each model. For each model, we must specify the likelihood function so that we can solve the hypothesis testing problem.

A complication, which arises in some cases, is that the sufficient statistic may not be monotonic. If it is monotonic, the decision regions Z_0 and Z_1 are simply connected: all portions of a region can be reached without crossing into the other region. If not, the regions are not simply connected and decision region islands are created. Disconnected regions usually complicate calculations of decision performance. Monotonic or not, the decision rule proceeds as described: the sufficient statistic is computed for each observation vector and compared to a threshold.

Example:
The coach of a soccer team suspects his goalie has been less than attentive to his training regimen. The coach focuses on the kicks the goalie makes to send the ball down the field. The data $\boldsymbol{r}$ he observes is the length of a kick. The coach defines the models as

- $\mathcal{M}_0$: not maintaining a training regimen
- $\mathcal{M}_1$: is maintaining a training regimen

The conditional densities---models---of the kick length are shown in [link].



Conditional densities for the distribution of the lengths of soccer kicks assuming that the goalie has not attended to his training ($\mathcal{M}_0$) or did ($\mathcal{M}_1$) are shown in the top row. The lower portion depicts the likelihood ratio formed from these densities.

Based on knowledge of soccer player behavior, the coach assigns a priori probabilities of $\pi_0 = 1/4$ and $\pi_1 = 3/4$. The costs C_{ij} are chosen to reflect the coach's sensitivity to the goalies feelings: $C_{01} = 1 = C_{10}$ (an erroneous decision either way is given the same cost) and $C_{00} = 0 = C_{11}$. The likelihood ratio is plotted in [link] and the threshold value η, which is computed from the a priori probabilities and the costs to be $1/3$, is indicated. The calculations of this comparison can be simplified in an obvious way.

$$\frac{r}{50} \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} 1/3$$

or

$$r \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} 50/3 = 16.7$$

The multiplication by the factor of 50 is a simple illustration of the reduction of the likelihood ratio to a sufficient statistic. Based on the assigned costs and a priori probabilities, the optimum decision rule says the coach must assume that the student did not train if a kick is less than 16.7; if greater, the goalie is assumed to have trained despite producing an abysmally short kick such as 20. Note that as the densities given by each model overlap entirely: the possibility of making the wrong interpretation **always** haunts the coach. However, no other procedure will be better (produce a smaller Bayes' cost)!

Detection Performance Criteria

The criterion used in the previous section---minimize the average cost of an incorrect decision---may seem to be a contrived way of quantifying decisions. Well, often it is. For example, the Bayesian decision rule depends explicitly on the a priori probabilities. A rational method of assigning values to these---either by experiment or through true knowledge of the relative likelihood of each model---may be unreasonable. In this section, we develop alternative decision rules that try to respond to such objections. One essential point will emerge from these considerations: **the likelihood ratio persists as the core of optimal detectors as optimization criteria and problem complexity change**. Even criteria remote from performance error measures can result in the likelihood ratio test. Such an invariance does not occur often in signal processing and underlines the likelihood ratio test's importance.

Maximizing the Probability of a Correct Decision

As only one model can describe any given set of data (the models are mutually exclusive), the probability of being correct P_c for distinguishing two models is given by

$$P_c = \mathrm{Pr}[\mathrm{say} \mathscr{M}_0 \mathrm{when} \mathscr{M}_0 \mathrm{true}] + \mathrm{Pr}[\mathrm{say} \mathscr{M}_1 \mathrm{when} \mathscr{M}_1 \mathrm{true}]$$

We wish to determine the optimum decision region placement. Expressing the probability of being correct in terms of the likelihood functions $\mathrm{p}_{\boldsymbol{R}|\mathscr{M}_i}\ (\boldsymbol{r})$, the a priori probabilities and the decision regions, we have

$$P_c = \int \pi_0\ \mathrm{p}_{\boldsymbol{R}|\mathscr{M}_0}\ (\boldsymbol{r})\ \mathrm{d}\ \boldsymbol{r} + \int \pi_1\ \mathrm{p}_{\boldsymbol{R}|\mathscr{M}_1}\ (\boldsymbol{r})\ \mathrm{d}\ \boldsymbol{r}$$

We want to maximize P_c by selecting the decision regions Z_0 and Z_1. Mimicking the ideas of the previous section, we associate each value of $\boldsymbol{r}$ with the largest integral in the expression for P_c. Decision region Z_0, for example, is defined by the collection of values of $\boldsymbol{r}$ for which the first term is largest. As all of the quantities involved are non-negative, the decision rule maximizing the probability of a correct decision is

Note: Given $\boldsymbol{r}$, choose $\mathscr{M}_i$ for which the product $\pi_i \ \mathrm{p}_{\boldsymbol{R}|\mathscr{M}_i}\left(\boldsymbol{r}\right)$ is largest.

When we must select among more than two models, this result still applies (prove this for yourself). Simple manipulations lead to the likelihood ratio test when we must decide between two models.

$$\frac{\mathrm{p}_{\boldsymbol{R}|\mathscr{M}_1}\left(\boldsymbol{r}\right)}{\mathrm{p}_{\boldsymbol{R}|\mathscr{M}_0}\left(\boldsymbol{r}\right)} \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \frac{\pi_0}{\pi_1}$$

Note that if the Bayes' costs were chosen so that $C_{ii} = 0$ and $C_{ij} = C$, ($i \neq j$), the Bayes' cost and the maximum-probability-correct thresholds would be the same.

To evaluate the quality of the decision rule, we usually compute the **probability of error** P_e rather than the probability of being correct. This quantity can be expressed in terms of the observations, the likelihood ratio, and the sufficient statistic.
Equation:

$$\begin{aligned} P_e &= \pi_0 \int \mathrm{p}_{\boldsymbol{R}|\mathscr{M}_0}\left(\boldsymbol{r}\right) \mathrm{d}\,\boldsymbol{r} + \pi_1 \int \mathrm{p}_{\boldsymbol{R}|\mathscr{M}_1}\left(\boldsymbol{r}\right) \mathrm{d}\,\boldsymbol{r} \\ &= \pi_0 \int \mathrm{p}_{\boldsymbol{\Lambda}|\mathscr{M}_0}\left(\boldsymbol{\Lambda}\right) \mathrm{d}\,\boldsymbol{\Lambda} + \pi_1 \int \mathrm{p}_{\boldsymbol{\Lambda}|\mathscr{M}_1}\left(\boldsymbol{\Lambda}\right) \mathrm{d}\,\boldsymbol{\Lambda} \\ &= \pi_0 \int \mathrm{p}_{\boldsymbol{\Upsilon}|\mathscr{M}_0}\left(\boldsymbol{\Upsilon}\right) \mathrm{d}\,\boldsymbol{\Upsilon} + \pi_1 \int \mathrm{p}_{\boldsymbol{\Upsilon}|\mathscr{M}_1}\left(\boldsymbol{\Upsilon}\right) \mathrm{d}\,\boldsymbol{\Upsilon} \end{aligned}$$

These expressions point out that the likelihood ratio and the sufficient statistic can each be considered a function of the observations $\boldsymbol{r}$; hence, they are random variables and have probability densities for each model. When the likelihood ratio is non-monotonic, the first expression is most difficult to evaluate. When monotonic, the middle expression often proves to be the most difficult. No matter how it is calculated, **no other decision rule can yield a smaller probability of error**. This statement is obvious as we minimized the probability of error implicitly by maximizing the probability of being correct because $P_e = 1 - P_c$.

From a grander viewpoint, these expressions represent an achievable lower bound on performance (as assessed by the probability of error). Furthermore, this probability will be non-zero if the conditional densities overlap over some range of values of $\boldsymbol{r}$, such as occurred in the previous example. Within regions of overlap, the observed values are ambiguous: either model is consistent with the observations. Our "optimum" decision rule operates in such regions by selecting that model which is most likely (has the highest probability) of generating the measured data.

Neyman-Pearson Criterion

Situations occur frequently where assigning or measuring the a priori probabilities π_i is unreasonable. For example, just what is the a priori probability of a supernova occurring in any particular region of the sky? We clearly need a model evaluation procedure that can function without a priori probabilities. This kind of test results when the so-called **Neyman-Pearson criterion** is used to derive the decision rule.

Using nomenclature from radar, where model $\mathscr{M}_1$ represents the presence of a target and $\mathscr{M}_0$ its absence, the various types of correct and incorrect decisions have the following names.[footnote]

- **Detection Probability** we say it's there when it is;
  $P_D = \Pr[\text{say} \mathscr{M}_1 \mid \mathscr{M}_1 \text{true}]$
- **False-alarm Probability** we say it's there when it's not;
  $P_F = \Pr[\text{say} \mathscr{M}_1 \mid \mathscr{M}_0 \text{true}]$
- **Miss Probability** we say it's not there when it is;
  $P_M = \Pr[\text{say} \mathscr{M}_0 \mid \mathscr{M}_1 \text{true}]$

The remaining probability $\Pr[\text{say} \mathscr{M}_0 \mid \mathscr{M}_0 \text{true}]$ has historically been left nameless and equals $1 - P_F$. We should also note that the detection and miss probabilities are related by $P_M = 1 - P_D$. As these are conditional probabilities, they do not depend on the a priori probabilities. Furthermore, the two probabilities P_F and P_D characterize the errors when **any** decision rule is used.
In statistics, a false-alarm is known as a **type I error** and a miss a **type II error**.

These two probabilities are related to each other in an interesting way. Expressing these quantities in terms of the decision regions and the likelihood functions, we have

$$P_F = \int \mathrm{p}_{\boldsymbol{R}|\mathscr{M}_0}(\boldsymbol{r})\,\mathrm{d}\,\boldsymbol{r}$$

$$P_D = \int \mathrm{p}_{\boldsymbol{R}|\mathscr{M}_1}(\boldsymbol{r})\,\mathrm{d}\,\boldsymbol{r}$$

As the region Z_1 shrinks, **both** of these probabilities tend toward zero; as Z_1 expands to engulf the entire range of observation values, they both tend toward unity. This rather direct relationship between P_D and P_F does not mean that they equal each other; in most cases, as Z_1 expands, P_D increases more rapidly than P_F (we had better be right more often than we are wrong!). However, the "ultimate" situation where a rule is always right and never wrong ($P_D = 1$, $P_F = 0$) cannot occur when the conditional distributions overlap. Thus, to increase the detection probability we must also allow the false-alarm probability to increase. **This behavior represents the fundamental tradeoff in detection theory**.

One can attempt to impose a performance criterion that depends only on these probabilities with the consequent decision rule not depending on the a priori probabilities. The Neyman-Pearson criterion assumes that the false-alarm probability is constrained to be less than or equal to a specified value α while we maximize the detection probability P_D.

$$\forall P_F, P_F \leq \alpha : (\max_{Z_1}\{Z_1, P_D\})$$

A subtlety of the solution we are about to obtain is that the underlying probability distribution functions may not be continuous, with the consequence that P_F can never equal the constraining value α. Furthermore, a (unlikely) possibility is that the optimum value for the false-alarm probability is somewhat less than the criterion value. Assume, therefore, that we rephrase the optimization problem by requiring that the false-alarm probability equal a value α' that is the largest possible value less than or equal to α.

This optimization problem can be solved using Lagrange multipliers; we seek to find the decision rule that maximizes

$$F = P_D - \lambda \left(P_F - \alpha' \right)$$

where λ is a positive Lagrange multiplier. This optimization technique amounts to finding the decision rule that maximizes F, then finding the value of the multiplier that allows the criterion toinge the detection probability in competition with false-alrm probabilities in excess of the criterion value. As is usual in the derivation of optimum decision rules, we maximize these quantities with respect to the decision regions. Expressing P_D and P_F in terms of them, we have
Equation:

$$\begin{aligned} F &= \int \mathrm{p}_{\boldsymbol{R}|\mathscr{M}_1}(\boldsymbol{r}) \, \mathrm{d}\, \boldsymbol{r} - \lambda \left(\int \mathrm{p}_{\boldsymbol{R}|\mathscr{M}_0}(\boldsymbol{r}) \, \mathrm{d}\, \boldsymbol{r} - \alpha' \right) \\ &= \lambda\alpha' + \int \left(\mathrm{p}_{\boldsymbol{R}|\mathscr{M}_1}(\boldsymbol{r}) - \lambda \, \mathrm{p}_{\boldsymbol{R}|\mathscr{M}_0}(\boldsymbol{r}) \right) \mathrm{d}\, \boldsymbol{r} \end{aligned}$$

To maximize this quantity with respect to Z_1, we need only to integrate over those regions of $\boldsymbol{r}$ where the integrand is positive). The region Z_1 thus corresponds to those values of $\boldsymbol{r}$ where $\mathrm{p}_{\boldsymbol{R}|\mathscr{M}_1}(\boldsymbol{r}) > \lambda \, \mathrm{p}_{\boldsymbol{R}|\mathscr{M}_0}(\boldsymbol{r})$ and the resulting decision rule is

$$\frac{\mathrm{p}_{\boldsymbol{R}|\mathscr{M}_1}(\boldsymbol{r})}{\mathrm{p}_{\boldsymbol{R}|\mathscr{M}_0}(\boldsymbol{r})} \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \lambda$$

The ubiquitous likelihood ratio test again appears; it **is** indeed the fundamental quantity in hypothesis testing. Using either the logarithm of the likelihood ratio or the sufficient statistic, this result can be expressed as

$$\ln(\Lambda(\boldsymbol{r})) \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \ln(\lambda)$$

or

$$\Upsilon(\boldsymbol{r}) \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \gamma$$

We have not as yet found a value for the threshold. The false-alarm probability can be expressed in terms of the Neyman-Pearson threshold in two (useful) ways.

Equation:

$$\begin{aligned} P_F &= \int_{\lambda}^{\infty} \mathrm{p}_{\boldsymbol{\Lambda}|\mathscr{M}_0}(\boldsymbol{\Lambda})\,\mathrm{d}\,\Lambda \\ &= \int_{\gamma}^{\infty} \mathrm{p}_{\boldsymbol{\Upsilon}|\mathscr{M}_0}(\boldsymbol{\Upsilon})\,\mathrm{d}\,\Upsilon \end{aligned}$$

One of these implicit equations must be solved for the threshold by setting P_F equal to α'. The selection of which to use is usually based on pragmatic considerations: the easiest to compute. From the previous discussion of the relationship between the detection and false-alarm probabilities, we find that to maximize P_D we must allow α' to be as large as possible while remaining less than α. Thus, we want to find the **smallest** value of λ consistent with the constraint. Computation of the threshold is problem-dependent, but a solution always exists.

Example:

An important application of the likelihood ratio test occurs when $\boldsymbol{R}$ is a Gaussian random vector for each model. Suppose the models correspond to Gaussian random vectors having different mean values but sharing the same covariance.

- $\mathscr{M}_0$: $\boldsymbol{R} \sim N\left(0, \sigma^2 I\right)$
- $\mathscr{M}_1$: $\boldsymbol{R} \sim N\left(\boldsymbol{m}, \sigma^2 I\right)$

$\boldsymbol{R}$ is of dimension L and has statistically independent, equi-variance components. The vector of means $\boldsymbol{m} = \left(m_0 \ldots m_{L-1}\right)^T$ distinguishes the two models. The likelihood functions associated this problem are

$$\mathrm{p}_{\boldsymbol{R}|\mathscr{M}_0}(\boldsymbol{r}) = \prod_{l=0}^{L-1} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\left(1/2\left(\frac{r_l}{\sigma}\right)^2\right)}$$

$$\mathrm{p}_{\boldsymbol{R}|\mathscr{M}_1}(\boldsymbol{r}) = \prod_{l=0}^{L-1} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\left(1/2\left(\frac{r_l-m_l}{\sigma}\right)^2\right)}$$

The likelihood ratio $\Lambda(\boldsymbol{r})$ becomes

$$\Lambda(\boldsymbol{r}) = \frac{\prod_{l=0}^{L-1} e^{-\left(1/2\left(\frac{r_l-m_l}{\sigma}\right)^2\right)}}{\prod_{l=0}^{L-1} e^{-\left(1/2\left(\frac{r_l}{\sigma}\right)^2\right)}}$$

This expression for the likelihood ratio is complicated. In the Gaussian case (and many others), we use the logarithm the reduce the complexity of the likelihood ratio and form a sufficient statistic.
Equation:

$$\begin{aligned} \ln(\Lambda(\boldsymbol{r})) &= \sum_{l=0}^{L-1} -1/2\frac{(r_l-m_l)^2}{\sigma^2} + 1/2\frac{{r_l}^2}{\sigma^2} \\ &= \frac{1}{\sigma^2}\sum_{l=0}^{L-1} m_l r_l - \frac{1}{2\sigma^2}\sum_{l=0}^{L-1} {m_l}^2 \end{aligned}$$

The likelihood ratio test then has the much simpler, but equivalent form

$$\sum_{l=0}^{L-1} m_l r_l \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \sigma^2 \ln(\eta) + 1/2\sum_{l=0}^{L-1} {m_l}^2$$

To focus on the model evaluation aspects of this problem, let's assume the means equal each other and are a positive constant: $m_l = m > 0$.
[footnote] We now have

$$\sum_{l=0}^{L-1} r_l \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \frac{\sigma^2}{m} \ln(\eta) + \frac{Lm}{2}$$

Note that all that need be known about the observations $\{r_l\}$ is their sum. This quantity is the sufficient statistic for the Gaussian problem: $\Upsilon(\boldsymbol{r}) = \sum r_l$ and $\gamma = \sigma^2 \ln\left(\frac{\eta}{m}\right) + \frac{Lm}{2}$.

What would happen if the mean were negative?

When trying to compute the probability of error or the threshold in the Neyman-Pearson criterion, we must find the conditional probability density of one of the decision statistics: the likelihood ratio, the log-likelihood, or the sufficient statistic. The log-likelihood and the sufficient statistic are quite similar in this problem, but clearly we should use the latter. One practical property of the sufficient statistic is that it usually simplifies computations. For this Gaussian example, the sufficient statistic is a Gaussian random variable under each model.

- $\mathscr{M}_0$: $\Upsilon(\boldsymbol{r}) \sim N\left(0, L\sigma^2\right)$
- $\mathscr{M}_1$: $\Upsilon(\boldsymbol{r}) \sim N\left(Lm, L\sigma^2\right)$

To find the probability of error from [link], we must evaluate the area under a Gaussian probability density function. These integrals are succinctly expressed in terms of $Q(x)$, which denotes the probability that a unit-variance, zero-mean Gaussian random variable exceeds x.

$$Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\alpha^2}{2}} \, \mathrm{d}\, \alpha$$

As $1 - Q(x) = Q(-x)$, the probability of error can be written as

$$P_e = \pi_1 Q\left(\frac{Lm - \gamma}{\sqrt{L}\sigma}\right) + \pi_0 Q\left(\frac{\gamma}{\sqrt{L}\sigma}\right)$$

An interesting special case occurs when $\pi_0 = 1/2 = \pi_1$. In this case, $\gamma = \frac{Lm}{2}$ and the probability of error becomes

$$P_e = Q\left(\frac{\sqrt{L}m}{2\sigma}\right)$$

As $Q(\cdot)$ is a monotonically decreasing function, the probability of error decreases with increasing values of the ratio $\frac{\sqrt{L}m}{2\sigma}$. However, as shown in [link], $Q(\cdot)$ decreases in a nonlinear fashion. Thus, increasing m by a factor of two may decrease the probability of error by a larger **or** a smaller factor; the amount of change depends on the initial value of the ratio.
To find the threshold for the Neyman-Pearson test from the expressions given on [link], we need the area under a Gaussian density.
Equation:

$$\begin{aligned} P_F &= Q\left(\frac{\gamma}{\sqrt{L\sigma^2}}\right) \\ &= \alpha' \end{aligned}$$

As $Q(\cdot)$ is a monotonic and continuous function, we can set α' equal to the criterion value α with the result

$$\gamma = \sqrt{L}\sigma Q^{-1}(\alpha)$$

where $Q^{-1}(\cdot)$ denotes the inverse function of $Q(\cdot)$. The solution of this equation cannot be performed analytically as no closed form expression exists for $Q(\cdot)$ (much less its inverse function). The criterion value must be found from tables or numerical routines. Because Gaussian problems arise frequently, the accompanying table provides numeric values for this quantity at the decade points.

x	$Q^{-1}(x)$
10^{-1}	1.281
10^{-2}	2.396

x	$Q^{-1}(x)$
10^{-3}	3.090
10^{-4}	3.719
10^{-5}	4.265
10^{-6}	4.754

The table displays interesting values for $Q^{-1}(\cdot)$ that can be used to determine thresholds in the Neyman-Pearson variant of the likelihood ratio test. Note how little the inverse function changes for decade changes in its argument; $Q(\cdot)$ is indeed **very** nonlinear. The detection probability of the Neyman-Pearson decision rule is given by

$$P_D = Q\left(Q^{-1}(\alpha) - \frac{\sqrt{L}m}{\sigma}\right)$$

Detection of Signals in Noise
Detection theory is specialized to the most common decision problem that occurs in signal processing: determining which signal was received in the presence of additive noise.

Far and away the most common decision problem in signal processing is determining which of several signals occurs in data contaminated by additive noise. Specializing to the case when one of two possible of signals is present, the data models are

- $\mathscr{M}_0 : R(l) = s_0(l) + N(l), 0 \leq l < L$
- $\mathscr{M}_1 : R(l) = s_1(l) + N(l), 0 \leq l < L$

where $\{s_i(l)\}$ denotes the known signals and $N(l)$ denotes additive noise modeled as a stationary stochastic process. This situation is known as the **binary detection problem**: distinguish between two possible signals present in a noisy waveform.

We form the discrete-time observations into a vector: $\boldsymbol{R} = (R(0) \ldots R(L-1))^T$. Now the models become

- $\mathscr{M}_0 : \boldsymbol{R} = \boldsymbol{s}_0 + \boldsymbol{N}$
- $\mathscr{M}_1 : \boldsymbol{R} = \boldsymbol{s}_1 + \boldsymbol{N}$

To apply our detection theory results, we need the probability density of $\boldsymbol{R}$ under each model. As the only probabilistic component of the observations is the noise, the required density for the detection problem is given by

$$\mathrm{p}_{\boldsymbol{R}|\mathscr{M}_i}(\boldsymbol{r}) = \mathrm{p}_{\boldsymbol{N}}(\boldsymbol{r} - \boldsymbol{s}_i)$$

and the corresponding likelihood ratio by

$$\Lambda(\boldsymbol{r}) = \frac{\mathrm{p}_{\boldsymbol{N}}(\boldsymbol{r} - \boldsymbol{s}_1)}{\mathrm{p}_{\boldsymbol{N}}(\boldsymbol{r} - \boldsymbol{s}_0)}$$

Much of detection theory revolves about interpreting this likelihood ratio and deriving the detection threshold.

Additive White Gaussian Noise

By far the easiest detection problem to solve occurs when the noise vector consists of statistically independent, identically distributed, Gaussian random variables, what is commonly termed **white Gaussian noise**. The mean of white noise is usually taken to be zero[footnote] and each component's variance is σ^2. The equal-variance assumption implies the noise characteristics are unchanging throughout the entire set of observations. The probability density of the noise vector evaluated at $\boldsymbol{r} - \boldsymbol{s}_i$ equals that of a Gaussian random vector having independent components with mean $\boldsymbol{s}_i$.

$$\mathrm{p}_{\boldsymbol{N}}\left(\boldsymbol{r} - \boldsymbol{s}_i\right) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{L}{2}} e^{-\left(\frac{1}{2\sigma^2}\left(\boldsymbol{r}-\boldsymbol{s}_i\right)^T\left(\boldsymbol{r}-\boldsymbol{s}_i\right)\right)}$$

The resulting detection problem is similar to the Gaussian example we previously examined, with the difference here being a non-zero mean---the signal---under both models. The logarithm of the likelihood ratio becomes

$$\left(\boldsymbol{r} - \boldsymbol{s}_0\right)^T\left(\boldsymbol{r} - \boldsymbol{s}_0\right) - \left(\boldsymbol{r} - \boldsymbol{s}_1\right)^T\left(\boldsymbol{r} - \boldsymbol{s}_1\right) \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} 2\sigma^2 \ln(\eta)$$

The usual simplifications yield in

$$\boldsymbol{r}^T\boldsymbol{s}_1 - \frac{{\boldsymbol{s}_1}^T\boldsymbol{s}_1}{2} - \left(\boldsymbol{r}^T\boldsymbol{s}_0 - \frac{{\boldsymbol{s}_0}^T\boldsymbol{s}_0}{2}\right) \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \sigma^2 \ln(\eta)$$

The model-specific components on the left side express the signal processing operations for each model.[footnote]
The zero-mean assumption is realistic for the detection problem. If the mean were non-zero, simply subtracting it from the observed sequence results in a zero-mean noise component.
If more than two signals were assumed possible, quantities such as these would need to be computed for each signal and the largest selected.

Each term in the computations for the optimum detector has a signal processing interpretation. When expanded, the term $\boldsymbol{s}_i{}^T \boldsymbol{s}_i$ equals $\sum_{l=0}^{L-1} {s_i}^2(l)$, the **signal energy** E_i. The remaining term, $\boldsymbol{r}^T \boldsymbol{s}_i$, is the only one involving the observations and hence constitutes the sufficient statistic $\Upsilon_i(\boldsymbol{r})$ for the additive white Gaussian noise detection problem.

$$\Upsilon_i(\boldsymbol{r}) = \boldsymbol{r}^T \boldsymbol{s}_i$$

An abstract, but physically relevant, interpretation of this important quantity comes from the theory of linear vector spaces. In that context, the quantity $\boldsymbol{r}^T \boldsymbol{s}_i$ would be termed the **projection** of $\boldsymbol{r}$ onto $\boldsymbol{s}_i$. From the Schwarz inequality, we know that the largest value of this projection occurs when these vectors are proportional to each other. Thus, a projection measures how much alike two vectors are: they are completely alike when they are parallel (proportional to each other) and completely dissimilar when orthogonal (the projection is zero). In effect, the projection operation removes those components from the observations which are orthogonal to the signal, thereby generalizing the familiar notion of filtering a signal contaminated by broadband noise. In filtering, the signal-to-noise ratio of a bandlimited signal can be drastically improved by lowpass filtering; the output would consist only of the signal and "in-band" noise. The projection serves a similar role, ideally removing those "out-of-band" components (the orthogonal ones) and retaining the "in-band" ones (those parallel to the signal).

Matched Filtering

The projection operation can be expanded as $\boldsymbol{r}^T \boldsymbol{s}_i = \sum_{l=0}^{L-1} r(l) s_i(l)$ another signal processing interpretation emerges. The projection now describes a finite impulse response (FIR) filtering operation evaluated at a specific index. To demonstrate this interpretation, let $h(l)$ be the unit-sample response of a linear, shift-invariant filter where $h(l) = 0$ for $l < 0$ and $l \geq L$. Letting $r(l)$ be the filter's input sequence, the convolution sum expresses the output.

$$r(k)*h(k) = \sum_{l=k-(L-1)}^{k} r(l)h(k-l)$$

Letting $k = L-1$, the index at which the unit-sample response's last value overlaps the input's value at the origin, we have

$$r(k)*h(k)|_{k=L-1} = \sum_{l=0}^{L-1} r(l)h(L-1-l)$$

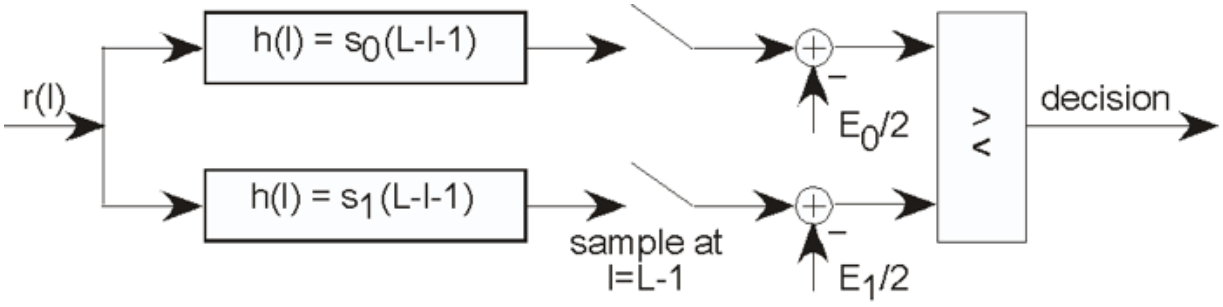
Suppose we set the unit-sample response equal to the index-reversed, then delayed signal.

$$h(l) = s_i(L-1-l)$$

In this case, the filtering operation becomes a projection operation.

$$r(k)*s_i(L-1-k)|_{k=L-1} = \sum_{l=0}^{L-1} r(l)s_i(l)$$

[link] depicts these computations graphically.



The detector for signals contained in additive, white Gaussian noise consists of a matched filter, whose output is sampled at the duration of the signal and half of the signal energy is subtracted from it. The optimum detector incorporates a matched filter for each signal compares their outputs to determine the largest.

The sufficient statistic for the i^{th} signal is thus expressed in signal processing notation as $r(k)^* s_i(L-1-k)|_{k=L-1} - \frac{E_i}{2}$. The filtering term is called a **matched filter** because the observations are passed through a filter whose unit-sample response "matches" that of the signal being sought. We sample the matched filter's output at the precise moment when all of the observations fall within the filter's memory and then adjust this value by half the signal energy. The adjusted values for the two assumed signals are subtracted and compared to a threshold.

Detection Performance

To compute the performance probabilities, the expressions should be simplified in the ways discussed in previous sections. As the energy terms are known a priori they can be incorporated into the threshold with the result

$$\sum_{l=0}^{L-1} r(l)\left(s_1(l) - s_0(l)\right) \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \sigma^2 \ln(\eta) + \frac{E_1 - E_0}{2}$$

The left term constitutes the sufficient statistic for the binary detection problem. Because the additive noise is presumed Gaussian, the sufficient statistic is a Gaussian random variable no matter which model is assumed. Under $\mathscr{M}_i$, the specifics of this probability distribution are

$$\sum_{l=0}^{L-1} r(l)\left(s_1(l) - s_0(l)\right) \sim N(m_i, \operatorname{var}_i)$$

where the mean and variance of the Gaussian distribution are given respectively by

$$m_i = \sum s_i(l)\left(s_1(l) - s_0(l)\right)$$

$$\mathrm{var}_i = \sigma^2 \sum \left(s_1(l) - s_0(l) \right)^2$$

Note that the variance does not depend on model. The false-alarm probability is given by

$$P_F = Q\left(\frac{\sigma^2 \ln(\eta) + \frac{E_1 - E_0}{2} - m_0}{\mathrm{var}^{\frac{1}{2}}} \right)$$

The signal-related terms in the numerator of this expression can be manipulated so that the false-alarm probability of the optimal white Gaussian noise detector is succinctly expressed by

$$P_F = Q\left(\frac{\ln(\eta) + \frac{1}{2\sigma^2} \sum \left(s_1(l) - s_0(l) \right)^2}{\frac{1}{\sigma} \left(\sum \left(s_1(l) - s_0(l) \right)^2 \right)^{\frac{1}{2}}} \right)$$

Note that the **only** signal-related quantity affecting this performance probability (and all of the others as well) is the **ratio of the energy in the difference signal to the noise variance.** The larger this ratio, the better (i.e., smaller) the performance probabilities become. Note that the details of the signal waveforms do not greatly affect the energy of the difference signal. For example, consider the case where the two signal energies are equal ($E_0 = E_1 = E$); the energy of the difference signal is given by $2E - 2 \sum s_0(l) s_1(l)$. The largest value of this energy occurs when the signals are negatives of each other, with the difference-signal energy equaling $4E$. Thus, equal-energy but opposite-signed signals such as sine waves, square-waves, Bessel functions, etc. **all** yield exactly the same performance levels. The essential signal properties that do yield good performance values are elucidated by an alternate interpretation. The term $\sum \left(s_1(l) - s_0(l) \right)^2$ equals $\left(\parallel s_1 - s_0 \parallel \right)^2$, the L^2 norm of the difference signal. Geometrically, the difference-signal energy is the same quantity as the square of the Euclidean distance between the two signals. In these terms, a larger distance between the two signals means better performance.

Example:
Detection, Gaussian example
A common detection problem is to determine whether a signal is present ($\mathscr{M}_1$) or not ($\mathscr{M}_0$). To model the latter case, the signal equals zero: $s_0(l) = 0$. The optimal detector relies on filtering the data with a matched filter having a unit-sample response based on the signal that might be present. Letting the signal under $\mathscr{M}_1$ be denoted simply by $s(l)$, the optimal detector consists of

$$r(l)*s(L-1-l)|_{l=L-1} - \frac{E}{2} \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \sigma^2 \ln(\eta)$$

or

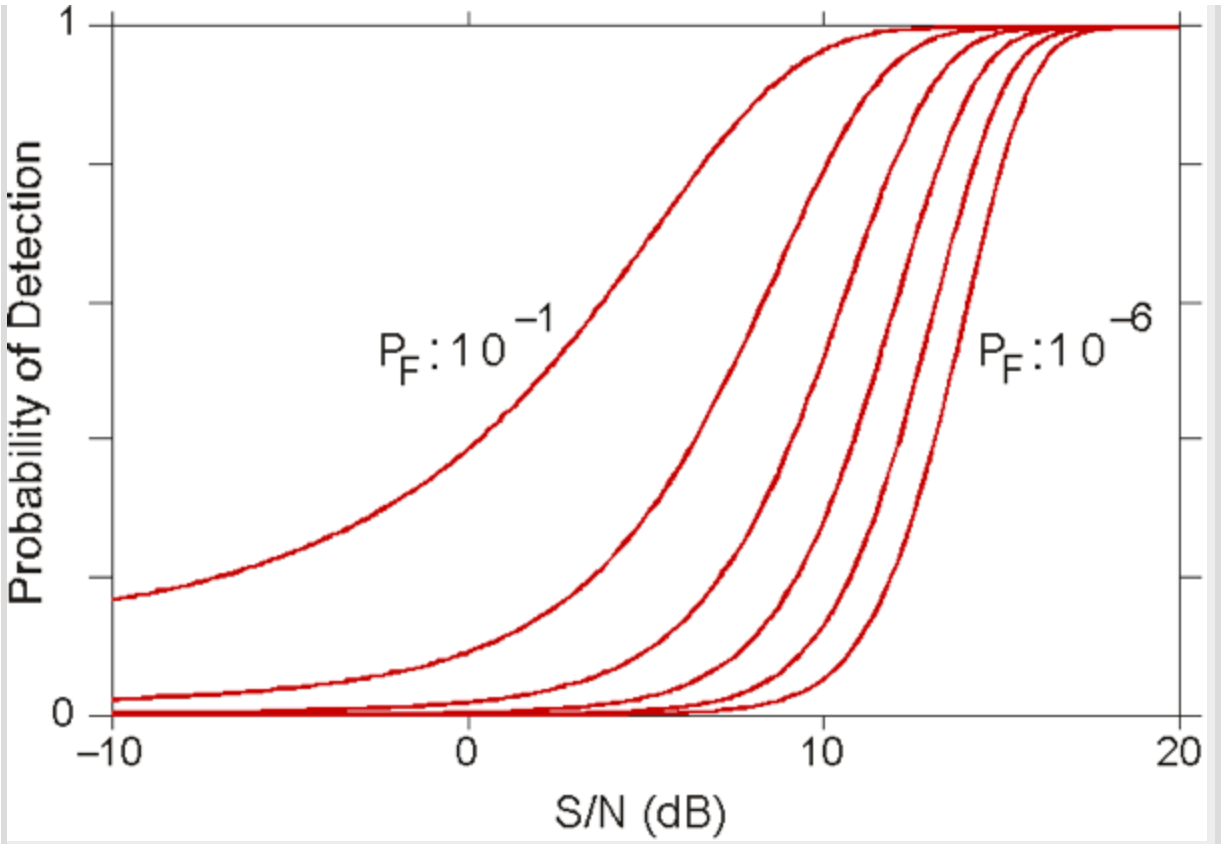
$$r(l)*s(L-1-l)|_{l=L-1} \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \gamma$$

The false-alarm and detection probabilities are given by

$$P_F = Q\left(\frac{\gamma}{\frac{E^{\frac{1}{2}}}{\sigma}}\right)$$

$$P_D = Q\left(Q^{-1}(P_F) - \sqrt{\frac{E}{\sigma}}\right)$$

[link] displays the probability of detection as a function of the signal-to-noise ratio $\frac{E}{\sigma^2}$ for several values of false-alarm probability. Given an estimate of the expected signal-to-noise ratio, these curves can be used to assess the trade-off between the false-alarm and detection probabilities.

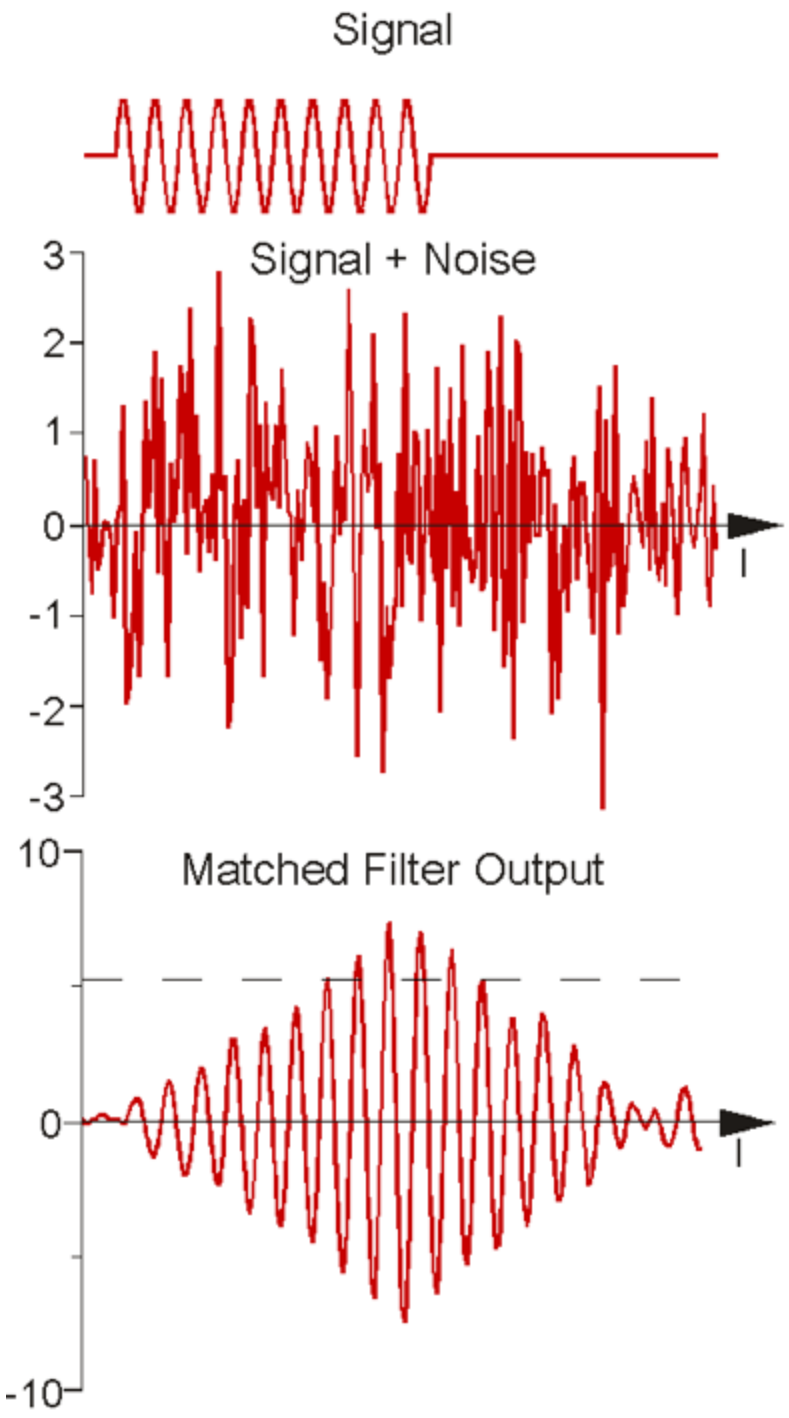


The probability of detection is plotted versus signal-to-noise ratio for various values of the false-alarm probability P_F. False-alarm probabilities range from 10^{-1} down to 10^{-6} by decades. The matched filter receiver was used since the noise is white and Gaussian. Note how the range of signal-to-noise ratios over which the detection probability changes shrinks as the false-alarm probability decreases. This effect is a consequence of the non-linear nature of the function $Q(\cdot)$.

The important parameter determining detector performance derived in this example is the **signal-to-noise ratio** $\frac{E}{\sigma^2}$: the larger it is, the smaller the false-alarm probability is (generally speaking). Signal-to-noise ratios can be measured in many different ways. For example, one measure might be the ratio of the rms signal amplitude to the rms noise amplitude. Note that the

important one for the detection problem is much different. The signal portion is the **sum** of the squared signal values over the **entire** set of observed values - the signal energy; the noise portion is the variance of **each** noise component - the noise power. Thus, energy can be increased in two ways that increase the signal-to-noise ratio: the signal can be made larger **or** the observations can be extended to encompass a larger number of values.

To illustrate this point, how a matched filter operates is shown in [link]. The signal is very difficult to discern in the presence of noise. However, the signal-to-noise ratio that determines detection performance belies the eye. The matched filter output demonstrates an amazingly clean signal.



The signal consists of ten cycles of $\sin(\omega_0 l)$ with $\omega_0 = 2\pi 0.1$. The middle panel shows the signal with noise added. The lower portion depicts the matched-filter output. The detection threshold was

set for a false-alarm probability of 10^{-2}. Even though the matched filter output crosses the threshold several times, only the output at $l = L - 1$ matters. For this example, it coincides with the peak output of the matched filter.

For more...

Many problems in statistical signal processing and communications can be solved using basic detection theory. For example, determining whether an airplane is located at a specific range and direction with radar and whether a received bit is a 0 or a 1 are both solved with matched filter detectors. However, many elaborations of deciding which of two models best describes a given dataset abound. For instance, suppose we have more than two models for the data. Previous modules hint at how to expand beyond two models (see [link] and [link]). However, no extensions for Neyman-Pearson detectors for more than two models exists.

To learn more about the basics of detection theory and beyond, see the books by Van Trees, Kay and McDonough and Whalen, and several modules on Connexions (search for 'likelihood ratio', 'detection theory' and 'matched filter'). The Wikipedia article on Statistical Hypothesis Testing describes what is called detection theory here more abstractly from a statistician's viewpoint.

Beyond Simple Problems

More interesting (and challenging) are situations where the data models are imprecise to some degree. The simplest case is when some model parameter is not known. For example, suppose the exact time of the radar return is not known (i.e., the airplane's range is uncertain). Here, the unknown parameter is the signal's time-of-origin. We must somehow determine that parameter **and** determine if the signal is actually present.

As you might expect, the likelihood ratio remains the focus of attention, now in the guise of what is known as the **generalized likelihood ratio test** (GLRT) (see this Connexions module).[footnote] This technique and others opens the door to what are known as **simultaneous estimation and detection algorithms**.
The Wikipedia article on the Likelihood-ratio test is concerned with the Generalized Likelihood Ratio Test.

Some unknowns may not be parametric and prevent a precise description of a model by a probability function. What do we do when the amplitude distribution function of the additive noise is not well characterized? So-called **robust detection** represents one attempt to address these problems. See [link] for more.

Beyond variations of model uncertainties are new approaches to solving detection problems. For example, a subtlety in the basic formulation is that all the data are available. Can we do just as well by attempting to make a decision prematurely as the data arrive? Note that always taking a **fixed** fewer number of samples always leads to worse performance. Instead, we acquire only the amount of data needed to make a decision. This approach is known as **sequential detection** or the **sequential probability ratio test** (SPRT). See modules in Connexions and the classic book by Wald.